

# Détection de fraudes et loi de probabilité de Newcomb-Benford

FERNANDEZ Christelle    PONCHEELE Clément    EL KAÏM Laura  
Encadré par M.DUCHARME

5 *mars* 2021

RESUME DU PROJET EN QLQ LIGNES

REMERCIEMENTS

## Table des matières

<b>Introduction.</b>	<b>2</b>
<b>#Tests</b>	<b>5</b>

# Introduction.

La fraude est une pratique répandue dans de nombreux domaines comme la finance, le secteur social ou médical. Il peut être tentant pour un être humain ou une société de tricher si cela peut impliquer pour lui une position plus confortable dans la société, telle qu'une réduction de charges, ou même un avantage sur un de ses concurrents. Il semblerait donc logique que des personnes cherchent à déceler ces fraudes.

Les données transmises par un individu ou un organisme peuvent faire l'objet de modifications, c'est de ce type de fraudes auquel nous nous intéresserons ici, et plus particulièrement la modification du premier chiffre significatif (le premier chiffre d'un nombre qui n'est pas un zéro) de nombres pris dans un certain ensemble de données.

De telles modifications entraînent un changement de la répartition des chiffres présents naturellement<sup>1</sup>. Si nous connaissons la répartition des chiffres présentés dans un ensemble de données arbitraires, il est donc techniquement possible de savoir si un nombre a été modifié ou non.

Il nous vient donc les questions suivantes : *Qu'elle est cette répartition ? Est-il possible de la connaître et si oui, dans quels cas ?*

De manière intuitive nous pourrions penser que les nombres sont répartis de manière uniforme. Qu'en est-il vraiment ?

La première partie de notre projet consistera à **répondre à ces questions**, nous nous appuyerons sur les travaux de Simon Newcom et Frank Benford, qui ont théorisé la **loi de Newcomb-Benford**, plus communément appelée loi de Benford. Cette loi nous dit que, dans une liste de données dites naturelles, la probabilité d'avoir le chiffre  $i$  comme premier chiffre significatif est de  $\log_{10}(1 + \frac{1}{i})$ .

Par exemple, le chiffre 1 en tant que premier chiffre significatif serait présent à hauteur de 30% alors que le 9 à seulement 4,6%.

Dans la suite **nous mettrons en œuvre une série d'expérimentation** pour constater ou non la véracité de cette loi, pour ce faire dans un premier temps nous récolterons des nombres pris dans des milieux sensés satisfaire la loi de Newcomb-Benford et observerons la répartition du premier chiffre significatif. Puis nous répliquerons une version simplifiée de l'expérience de Hill (1988), qui consiste à observer la répartition du premier chiffre significatif d'une liste de nombre donnée au hasard par des êtres humains, en l'occurrence ses élèves.

Cette expérience est à la base des méthodes de détection de fraudes par la loi de Newcomb-Benford. Si un fraudeur modifie un jeu de données, ce jeu est donc influencé par la pensée humaine, il ne suit donc plus la loi de Newcomb-Benford. Pour détecter la fraude il suffit donc de comparer les premiers chiffres significatifs. Cependant ces comparaisons doivent se faire de manière rigoureuses et scientifiques. Pour cela il existe des test statistiques, dont le plus connu, le test du  $\chi^2$ , ou bien celui de Ducharme et collab. (2020).

Il nous vient donc les questions suivantes : *Ces test sont-ils fiable ? Existe-t-il un test significativement meilleur que les autres ? Vont-ils dans le même sens ? Et sinon que faire ?*

---

1. Les données dites naturelles sont celles qui n'ont pas été influencé par la pensée de l'homme. .

La réponse à ces question constituera donc la deuxième partie de ce projet, pour ce faire nous mettrons en oeuvre différents tests sur des jeux de données comme la fiscalité italienne.

## #Les origines/Description

C'est la question que se posera en premier lieu l'astronome, mathématicien, économiste et statisticien canadien Simon Newcomb. En 1881, celui-ci fournira une première approche au principe statistique qui se fera appeler *Loi de Benford*.

Cette découverte mise de côté pendant plusieurs années, ce n'est qu'en 1938 que l'ingénieur et physicien américain Frank Benford pensera être le premier à l'initiative de cette loi. Les articles de Newcomb et Benford proposeront finalement les mêmes résultats. D'après leurs travaux, les nombres ne sont pas répartis de manière uniforme. On retrouve ce données dans énormément de domaine comme les mathématiques, l'environnement, la finance, la physique. Par exemples, la longueurs des fleuves, la population des villes dans un pays, les déclarations de revenus.

#Experimentations

##Benford

Intéressons-nous ici à la suite de Fibonacci.

Cette suite est une suite d'entiers dans laquelle chaque terme est la somme des deux termes qui le précèdent. Sa formulation est la suivante :

$$F_0 = 0, F_1 = 1, \text{ et } \forall n \geq 2, F_n = F_{n-1} + F_{n-2}.$$

Commençons par extraire les 300 premiers termes de la suite de Fibonacci.

Puis nous extrayons le premier chiffre significatif de chacun de ces nombres.

PCS	Proba	Benford
1	0.3033333	0.3010300
2	0.1733333	0.1760913
3	0.1266667	0.1249387
4	0.0900000	0.0969100
5	0.0833333	0.0791812
6	0.0633333	0.0669468
7	0.0566667	0.0579919
8	0.0566667	0.0511525
9	0.0433333	0.0457575

##Pas Benford

**#Tests**

#Bibliographie »»»> clement